PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases

Chenggang Yu, Nela Zavaljevski, Valmik Desai, and Jaques Reifman*

Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Fort Detrick, Maryland, USA

ABSTRACT

In this article, we present a new method termed CatFam (Catalytic Families) to automatically infer the functions of catalytic proteins, which account for 20–40% of all proteins in living organisms and play a critical role in a variety of biological processes. CatFam is a sequence-based method that generates sequence profiles to represent and infer protein catalytic functions. CatFam generates profiles through a stepwise procedure that carefully controls profile quality and employs nonenzymes as negative samples to establish profile-specific thresholds associated with a predefined nominal false-positive rate (FPR) of predictions. The adjustable FPR allows for fine precision control of each profile and enables the generation of profile databases that meet different needs: function annotation with high precision and hypothesis generation with moderate precision but better recall. Multiple tests of CatFam databases (generated with distinct nominal FPRs) against enzyme and nonenzyme datasets show that the method's predictions have consistently high precision and recall. For example, a 1% FPR database predicts protein catalytic functions for a dataset of enzymes and nonenzymes with 98.6% precision and 95.0% recall. Comparisons of CatFam databases against other established profile-based methods for the functional annotation of 13 bacterial genomes indicate that CatFam consistently achieves higher precision and (in most cases) higher recall, and that (on average) CatFam provides 21.9% additional catalytic functions not inferred by the other similarly reliable methods. These results strongly suggest that the proposed method provides a valuable contribution to the automated prediction of protein catalytic functions. The CatFam databases and the database search program are freely available at http://www.bhsai.org/downloads/catfam.tar.gz.

Proteins 2009; 74:449–460.

Key words: catalytic function; automated annotation; Enzyme Commission number; profile method; protein function prediction.

INTRODUCTION

The continual advancements in genome sequencing technology are contributing to the exponential increase of the rate at which we accumulate protein sequence data.[1,2] Unfortunately, our ability to experimentally ascertain the function and annotate protein sequences has not increased at the same rate, continually increasing the gap between protein sequence data and their functional annotation.[3,4] Hence, although not perfect, computational methods arguably offer the only feasible solution for addressing this disparity and providing high-throughput annotation of protein function. Among the various protein functions to be annotated, enzyme catalytic functions are of great importance because about 20–40% of the genes in genomes of the three domains of life code for enzymes,[5] which play many critical roles in a variety of biological processes in living organisms.[6–8]

Traditionally, the computational prediction of protein catalytic functions has been based on function transfer among homologous proteins, which assumes that functions are shared among proteins with similar sequences or structures.[9] BLAST[10] and other equivalent search methods have enabled for fast and efficient searches of similar sequences in large databases. It has become a common practice to perform BLAST searches of a query protein against a function-annotated protein sequence database, such as the Swiss-Prot database (http://expasy.org/sprot/), and transfer the annotated proteins' functions to the query protein for those proteins that exceed a specified sequence similarity threshold (i.e., an *E*-value cutoff). However, the accuracy of such methods is frequently questioned. Although most proteins with high sequence similarity very likely share similar functions, exceptions have been reported. Particularly for enzymes, small changes in key residues have shown to change protein function.[11,12] More accurate function

Grant sponsor: The US Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes (HSAI) initiative.
*Correspondence to: Jaques Reifman, BHSAI/MRMC, Attn: MCMR–TT, Building 363 Miller Drive, Fort Detrick, MD 21702-5012. E-mail: jaques.reifman@us.army.mil
Received 8 January 2008; Revised 18 April 2008; Accepted 2 June 2008
Published online 17 July 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22167

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Medical Research and Materiel Command,Biotechnology High Performance Computing Software Applications Institute,Telemedicine and Advanced Technology Research Center,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this article, we present a new method termed Cat- Fam (Catalytic Families) to automatically infer the functions of catalytic proteins, which account for 20? 40% of all proteins in living organisms and play a critical role in a variety of biological processes. Cat- Fam is a sequence-based method that generates sequence profiles to represent and infer protein catalytic functions. CatFam generates profiles through a stepwise procedure that carefully controls profile quality and employs nonenzymes as negative samples to establish profile-specific thresholds associated with a predefined nominal false-positive rate (FPR) of predictions. The adjustable FPR allows for fine precision control of each profile and enables the generation of profile databases that meet different needs: function annotation with high precision and hypothesis generation with moderate precision but better recall. Multiple tests of CatFam databases (generated with distinct nominal FPRs) against enzyme and nonenzyme datasets show that the method?s predictions have consistently high precision and recall. For example, a 1% FPR database predicts protein catalytic functions for a dataset of enzymes and nonenzymes with 98.6% precision and 95.0% recall. Comparisons of CatFam databases against other established profile-based methods for the functional annotation of 13 bacterial genomes indicate that CatFam consistently achieves higher precision and (in most cases) higher recall, and that (on average) CatFam provides 21.9% additional catalytic functions not inferred by the other similarly reliable methods. These results strongly suggest that the proposed method provides a valuable contribution to the automated prediction of protein catalytic functions. The CatFam databases and the database search program are freely available at http://www.bhsai.org/ downloads/catfam.tar.gz.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**12**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

predictions may be achieved by structure-based homology methods. For example, George *et al.*[11] proposed a method based on the presence of particular amino acid residues at a few active sites in the three-dimensional (3D) structure of a protein.[13] However, the lack of 3D structural information for the majority of the sequenced proteins significantly limits its application. Homology-based methods are also constrained by the limited number of homologous proteins that have been well annotated and by the difficulty in making reliable predictions for proteins with very low sequence similarity. Homology methods can completely fail to determine "orphan enzyme" activity, that is, a catalytic activity for which no sequence information is available,[14–16] or the activity of an "orphan gene" that has no detectable homologs in other organisms.[16]

An alternative approach that may complement homology-based methods is the one based on *ab initio* methods. They employ statistical and machine-learning-based techniques,[17–22] such as Bayesian classification, decision trees, association rules, neural networks,[18] and support vector machines,[19,20] to classify protein catalytic functions using various features derived from sequence and/or structure information of proteins with known functions. These features include sequence-related physicochemical properties,[18,19] such as polarity, hydrophobicity, Van der Waals volume, and glycosylation, as well as structure-related information, such as predicted secondary structure.[18,22] Despite the potential complementary benefit of utilizing these approaches when homology methods fail, the reported *ab initio* methods[17–20,22] only predict the first two digits of the four-digit Enzyme Commission (EC) number used for catalytic function characterization. Therefore, given the ever increasing amount and availability of protein sequence data, improved sequence-based homology methods currently provide the most practical computational solution for predicting catalytic function on a genome-wide scale.

In one such effort, a novel probabilistic method was proposed[23,24] to improve catalytic function predictions based on BLAST searches of a database of annotated enzymes. For a query enzyme, the method takes into account all BLAST search results and employs Bayesian statistics to determine the most probable EC number for the query enzyme. The method predicts enzyme functions with high precision but is limited to the cases where the query protein is known to be an enzyme. A more general approach that has shown to significantly improve the accuracy of sequence-based methods[25] is to generate sequence "profiles" to characterize the functions of similar protein sequences that share a common functional annotation. Recently, two methods based on sequence profiles, PRIAM[26] and EFICAz,[27,28] have been proposed for predicting protein catalytic functions and have proven to be highly accurate in estimating catalytic functions represented by EC number for a variety of enzymes.

PRIAM and EFICAz generate enzyme profiles by segregating enzymes with known EC numbers gathered from the Swiss-Prot database. Enzymes that share a common EC number are grouped together to generate one or more profiles to characterize the EC number (or function) of the proteins in the group. In PRIAM, the shortest sequence in an EC group is selected as the seed for PSI-BLAST[25] searches against the proteins in the group. The searches generate a sequence profile, which is represented in the form of a Position Specific Scoring Matrix (PSSM), and identify enzymes in the group that are similar to the profile. The identified enzymes are then removed from the group, and the process is sequentially repeated, each time generating a new profile until all enzymes in the group have been removed. For all profiles, PRIAM uses the same *E*-value cutoff to determine whether to transfer the function of the profile to the query protein. Conversely, rather than sequentially generating multiple profiles for each EC number, EFICAz first clusters enzymes by sequence similarity within an EC number group and then uses Hidden Markov Models[29] to generate profiles for each cluster of enzymes. This reduces the possibility of separating proteins with very similar sequences in the generation of multiple profiles. More importantly, EFICAz employs sequence identity and negative samples, that is, proteins associated with functions different from the considered function, to establish a specific cutoff for each profile. This can be effective in reducing excessive false predictions for particular EC numbers, whereas methods that employ a single cutoff for all profiles (e.g., PRIAM) can only assure an overall, average performance.

In this article, we present a new method termed CatFam (Catalytic Families) to automatically infer protein catalytic functions using sequence profiles. CatFam's profile generation procedure is similar to that of EFICAz. It uses a hierarchical clustering algorithm to cluster enzyme sequences and employs negative samples to generate profile-specific cutoffs that determine whether to transfer the function of the profile to the query protein. However, CatFam employs ClustalW[30] and PSI-BLAST to generate profiles in PSSM format and uses a stepwise procedure to control the quality of a profile during its generation. More importantly, unlike EFICAz, which uses sequence identity between the query protein and the sequences used to generate the profile to determine whether to transfer the function of the profile to the query protein, CatFam uses the raw score threshold (RST) of the profile itself. In contrast to sequence identity, the raw score of the sequence-profile alignment provides a direct measure of the similarity between the query protein and the enzyme profiles characterizing the catalytic functions. Furthermore, this direct measure obviates the need to maintain a sequence database of the enzymes used to generate the profiles, which is needed to compute sequence identity of the query protein. Moreover, because

RST is associated with predefined nominal false positive rates (FPRs), it enables the generation of distinct profile databases with different levels of precision and recall that are yet to be implemented by other prediction methods.

Next, we present the CatFam profile generation algorithm. Then, we assess the performance of the CatFam enzyme profile databases in various test cases by comparing them against BLAST and the two well-established profile-based methods, PRIAM and EFICAz, for protein catalytic function prediction. Finally, we conclude by summarizing the major features of CatFam and contrasting it against the two profile-based methods.

METHOD

Data preparation

We employ enzyme and nonenzyme protein data annotated in the Swiss-Prot database to construct datasets for the generation and testing of CatFam databases. The enzyme data consist of protein sequences and their corresponding EC numbers. The EC numbers are consistent with records in the Enzyme Nomenclature Database (http://www.expasy.ch/enzyme/), which cross-reference all enzymes in Swiss-Prot. We label proteins as nonenzymes by following a rule adapted from that used by EFICAz: a protein in Swiss-Prot is classified as a nonenzyme if no EC number, no enzyme keywords, and no words indicating less reliable function annotation, such as hypothetical and putative, are associated with the protein.

We assume that the manual annotations in the Swiss-Prot database provide an appropriate "gold standard" to train CatFam. Although errors inevitably exist in this database, a recent study indicates that most errors are due to under-annotation, that is, missed enzyme annotations, and a substantial number of such omissions will be corrected in the next Swiss-Prot release.[31] Moreover, the detrimental effect of sporadic annotations of wrong protein functions for a small number of enzymes can be reduced when they are merged with a large number of correctly annotated enzymes to generate a sequence profile. In combination with precision control during profile generation, this ensures that annotations performed by CatFam do not lead to over-predictions, which are the most detrimental type of errors propagated in databases.[32]

The primary dataset used in this study consists of 189,178 proteins (75,687 enzymes and 113,491 nonenzymes) from Swiss-Prot released in November 2006. About 90% of the enzymes and nonenzymes are randomly selected to form a training dataset D_{tr} to generate CatFam databases. The remaining proteins, 7600 enzymes and 11,349 nonenzymes, are set aside to form a testing enzyme dataset D_{enz1} and a testing nonenzyme dataset D_{nz}, respectively. Using the latest Swiss-Prot release (August 2007), we form a secondary testing enzyme dataset D_{enz2}, consisting of 8399 newly added enzymes. We use D_{enz2} as a surrogate to assess how well the CatFam databases can predict the catalytic functions of future, yet unannotated proteins.

Enzyme profile database generation

A sequence profile generated from protein sequences of a common function reveals the functionally conserved amino acid patterns of the sequences. Hence, a protein that matches such a profile can be annotated by the function associated with the profile. We generate profiles from enzymes that are annotated with the same EC number in the training dataset D_{tr}. For each EC number g, one or more profiles are generated by the following procedures:

a. Create a subset of enzymes D_g from D_{tr}, consisting of enzymes with EC number g.

b. Compute the sequence similarity between each pair of enzymes in D_g. This is performed by all-against-all BLAST searches, where sequence similarity is measured based on the E-value of the alignment of each pair of sequences.

c. Cluster enzymes by their sequence similarity, that is, E-value, using a hierarchical clustering algorithm.[33] Initially, each sequence forms a cluster. Then, we perform a pairwise search among all clusters and merge two clusters, C_i and C_j, which have the smallest cost function

$$F(C_i, C_j) = \max[E(a, b), \forall a \in C_i, \forall b \in C_j], \qquad (1)$$

into one cluster. Here, $E(a,b)$ denotes the E-value between protein sequences a and b in clusters C_i and C_j, respectively. Next, we sequentially continue this merging procedure until the cost function F exceeds a preset limit, at which point we have partitioned D_g into a total of K distinct clusters C_k, $k = 1,2,\ldots,K$. The proteins in each cluster are used to initialize a profile-generation set S_k for cluster C_k.

d. Generate one profile for cluster C_k. This is achieved by using ClustalW to perform multiple sequence alignment (MSA) for protein sequences in set S_k, followed by PSI-BLAST searches to generate a PSSM format profile $p_{k,m}$, $m = 1,2,\ldots,M$, where M is the total number of profiles used to represent cluster C_k.

e. Expand the set S_k by adding one additional sequence s from D_g. The expanded set is used to generate a new profile for C_k. The added sequence s is selected as the one that is most similar to all sequences already in S_k, that is, the sequence that has the smallest cost function

$$F(s, S_k) = \max[E(s, b), \forall b \in S_k]. \qquad (2)$$

This gradual addition of divergent sequences preserves the quality of the MSA used to generate the profile for C_k.

f. Return to Step (d) to generate another profile $p_{k,m}$ for C_k unless one of the two following conditions that terminate the expansion of the set S_k in Step (e) is met: (1) there are no remaining proteins in D_g, or (2) the MSA does not have a single fully conserved position. The second condition prevents the addition of proteins to S_k that are too divergent and would have significantly lowered the quality of the generated profile for C_k.

g. Select the best profile for cluster C_k. A series of profiles $p_{k,1}, p_{k,2}, \ldots, p_{k,M}$ are generated through iterations of Steps (d) and (e). For each profile, we first use PSI-BLAST to align all proteins in D_{tr} with that profile. Then, we rank order the proteins according to their raw score value and compute the FPR (i.e., the fraction of proteins with EC number other than g) associated with each score. Next, starting with the largest raw score, we search through the list of ranked proteins and identify the one associated with the largest FPR that passes a predefined threshold. The corresponding raw score is labeled the RST, which is used as the cutoff for the profile. Finally, for each profile, we compute the number of true positives (i.e., number of proteins with EC number g) associated with the corresponding RST and select as the best profile for cluster C_k the one with the largest number of true positives.

h. Return to Step (d) to generate a profile for another cluster C_k until the procedure is completed for all K clusters.

This procedure allows the user to define the nominal FPR for each EC number during the generation of the CatFam databases. Therefore, the user can select low FPRs to generate databases with highly accurate enzyme annotation or high FPRs to construct databases with high recall.

RESULTS

We assess the performance of the CatFam databases by comparing and contrasting them against well-established resources for predicting protein catalytic function, such as BLAST, PRIAM, and EFICAz. The availability of BLAST and PRIAM's source code allows us to comparatively assess CatFam's performance for the three customized testing datasets, D_{enz1}, D_{enz2}, and D_{nz}, discussed earlier. Conversely, due to the unavailability of the EFICAz code, comparisons with it are limited to precomputed enzyme functions available on its Web-site in September 2007 (http://cssb2.biology.gatech.edu/EFICAz/).

CatFam databases

To test the performance of the proposed enzyme-prediction algorithm, we construct four CatFam databases, consisting of enzyme profiles and their associated EC numbers and raw score thresholds. We construct each of the four databases to satisfy one nominal FPR (1, 5, 10, and 25%) specified during profile generation and test their ability to predict enzymes labeled by four-digit EC numbers. Table I lists the distribution of distinct EC numbers used in the development of the CatFam databases. Out of the 2220 distinct EC numbers in the Swiss-Prot database, 1885 (86%) are covered in the training dataset D_{tr}. Each of the 335 not covered EC numbers corresponds to only one enzyme in the Swiss-Prot database, making them unsuitable for profile generation. For a 1% FPR, CatFam generates 8080 profiles for 1653 EC numbers, comprising 88% of the EC numbers in D_{tr}. For the remaining 12%, profiles are not generated because of the insufficient number of training enzymes. For the other FPRs, CatFam generates similar number of profiles, covering a comparable amount of EC numbers. Because the testing dataset D_{enz1} is created by randomly selecting 10% of the enzymes for each EC number, no testing enzymes are selected for EC numbers associated with less than 10 enzymes. Thus, the testing dataset D_{enz1} only covers about half of the EC numbers in the CatFam databases. Interestingly, the dataset D_{enz2}, which contains newly annotated enzymes, has even fewer distinct EC numbers than D_{enz1}.

Table I
Distribution of EC Numbers Used in the Development of the CatFam Databases

Distinct EC numbers in the Swiss-Prot database[a]	2220
Distinct EC numbers in the training dataset D_{tr}	1885
Number of profiles in the CatFam database[b]	8080
Distinct EC numbers in the CatFam database[b]	1653
Distinct EC numbers in D_{enz1}	856
Distinct EC numbers in D_{enz2}	545

[a]Swiss-Prot database released in November 2006.
[b]These numbers correspond to the CatFam database with 1% false positive rate. The numbers for other CatFam databases are slightly larger.

Assessment of CatFam's performance

We first assess the capability of the CatFam databases to discriminate between enzymes and nonenzymes and compare their performance against BLAST searches on the training dataset D_{tr}. A query protein is labeled as an enzyme if an EC number is assigned to it by CatFam or if a BLAST search finds an enzyme in D_{tr} with E-value less a given cutoff. Figure 1 shows the combined results against the enzyme (D_{enz1}) and nonenzyme (D_{nz}) datasets. As expected, smaller E-value cutoffs in BLAST searches decrease the false identification of nonenzymes, while increasing the misidentification of enzymes. The results of the CatFam databases with decreasing FPRs yield a similar trend. However, when compared with BLAST, for a fixed number of misidentified enzymes, CatFam yields a much smaller number of falsely identified nonenzymes. This suggests that the CatFam profiles are effective in characterizing enzyme catalytic functions, effectively distinguishing enzymes and nonenzymes.

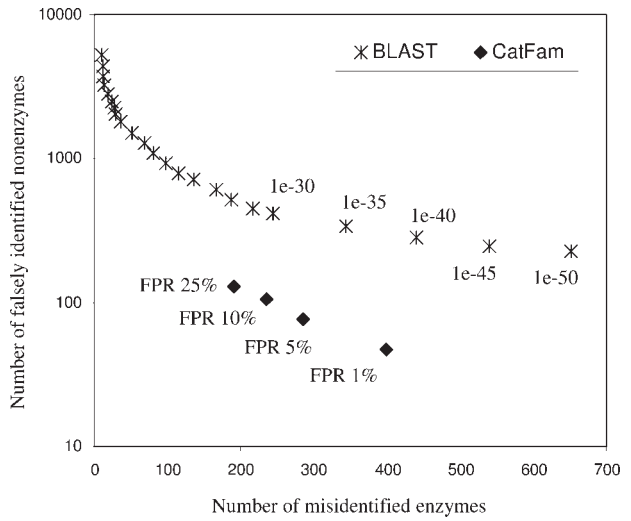

Figure 1

Comparison of CatFam databases and BLAST searches for the discrimination of enzymes and nonenzymes in datasets D_{enz1} and D_{nz}. A query protein is labeled as an enzyme if an EC number is assigned by CatFam or if a BLAST search against the training dataset D_{tr} finds an enzyme with *E*-value less than a given cutoff. The figure shows some of these *E*-value cutoffs.

We further assess the catalytic function prediction of the four CatFam databases by computing precision and recall for the two testing enzyme datasets, D_{enz1} and D_{enz2}, and for the nonenzyme dataset, D_{nz}, and comparing the results against PRIAM's predictions. Precision and recall are defined as:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$
$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. For each testing set, TP is the number of predicted EC numbers that are consistent with the proteins' original EC number assignments, FP is the number of predicted EC numbers that do not match the proteins' original assignments, and FN is the number of originally assigned EC numbers that are not predicted.

Table II compares the performance of the two methods for the different testing datasets. The results for the two testing enzyme datasets indicate each method's performance for the case where all proteins are enzymes. Situations like these may arise in genome reannotations performed to update or refine existing enzyme annotations. The results for the dataset that combines D_{enz1} and D_{nz} mimic the performance of automated protein annotations for newly sequenced genomes, involving both enzymes and nonenzymes.

As expected, the CatFam databases consistently achieve higher precision and lower recall with smaller preset FPRs. In the case of D_{enz1}, when the FPR changes from 25% to 1%, precision increases from 96.0 to 99.2% and recall decreases from 97.6 to 95.0%. When compared with D_{enz1}, the precision of each CatFam database for D_{enz2}, which consists of enzymes recently annotated and added to the Swiss-Prot database, drops by less than 1.0%. This suggests a consistently high reliability of CatFam's predictions. Conversely, recall for D_{enz2} drops slightly more than 10.0% for each of the CatFam databases. The lower recall is attributed to CatFam databases, trained on a previous release of Swiss-Prot, not being able to characterize new sequence patterns in enzymes recently added to the Swiss-Prot database. This is supported by our observation that there are as many as 665 (8%) proteins in D_{enz2}, compared with 51 (0.7%) proteins in D_{enz1}, that have less than 15% sequence similarity with the enzymes used to train the CatFam profiles, as shown in Figures 2(a,b). In addition, if enzymes in D_{enz2} have catalytic functions associated with orphan enzymes or orphan genes in the previous Swiss-Prot database, they will not be predicted either. When comparing the precision of the composite dataset that combines both enzymes and nonenzymes $D_{enz1} + D_{nz}$ with D_{enz1}, we find that the addition of nonenzymes causes a slight decrement in precision, monotonically decreasing it by 1.5–0.6% as the preset FPR changes from 25% to 1%. This further indicates that CatFam databases can accurately discriminate enzymes from nonenzymes.

Comparisons between PRIAM and the CatFam databases show that both PRIAM's precision and recall are consistently and systematically lower than those of all four CatFam databases' results for all testing datasets. Although PRIAM's precision for the enzyme datasets, D_{enz1} and D_{enz2}, is only about 4.0% lower than CatFam's results for 25% FPR, its recall is about 10.0% lower. Consistently, PRIAM's precision and recall for the composite dataset, $D_{enz1} + D_{nz}$, are about 10.0% lower than those for CatFam's with 25% FPR. These results clearly suggest that CatFam outperforms PRIAM in discriminating between enzymes and nonenzymes.

The performance of sequence-based protein function annotation methods is highly dependent on the sequence identity between the query protein and the proteins with known function used for method development. To

Table II

Comparison of Catalytic Function Predictions of Four CatFam Databases Versus PRIAM, Using Two Testing Enzyme Datasets, D_{enz1} and D_{enz2}, and One Nonenzyme Dataset, D_{nz}

Preset false positive rate (FPR)		CatFam 1%	CatFam 5%	CatFam 10%	CatFam 25%	PRIAM
D_{enz1}	Precision (%)	99.2	98.5	97.2	96.0	93.4
	Recall (%)	95.0	96.4	97.0	97.6	87.9
D_{enz2}	Precision (%)	99.0	97.9	96.6	95.3	91.4
	Recall (%)	82.3	84.5	86.3	87.4	76.3
$D_{enz1} + D_{nz}$	Precision (%)	98.6	97.5	95.9	94.5	82.6
	Recall (%)	95.0	96.4	97.0	97.6	87.9

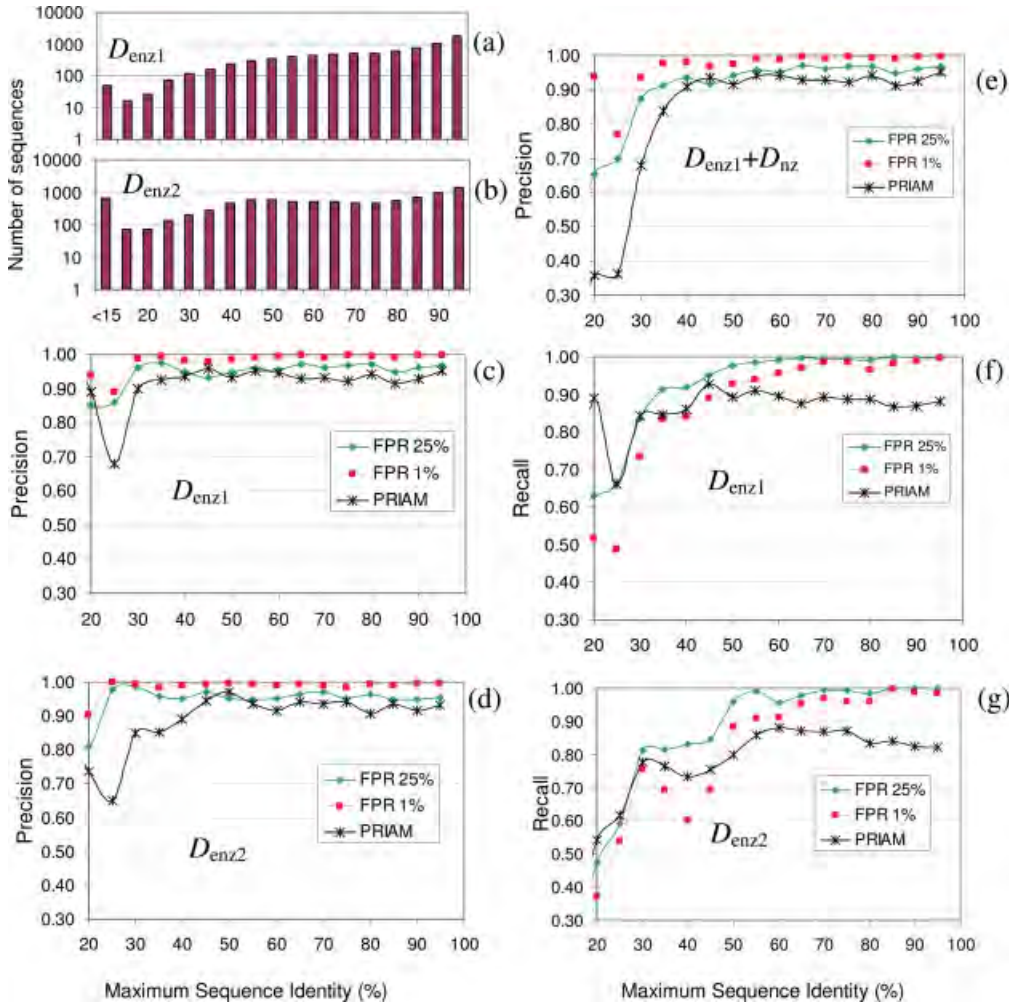

Figure 2

Precision and recall as a function of maximum sequence identity (MSI) for PRIAM and two catalytic family (CatFam) databases, one with 1% false positive rate (FPR) and another with 25% FPR. The MSI distribution for proteins in the testing datasets D_{enz1} and D_{enz2} are shown in Figure 2(a,b), respectively.

compare the performance of CatFam and PRIAM as a function of protein sequence identity, we sort the testing results according to the maximum sequence identity (MSI) between the query proteins and the proteins used for profile generation. In such comparison, we make an assumption that the distribution of proteins used to construct PRIAM, which we do not know, would result in similar MSI when queried against D_{enz1} and D_{enz2}. This assumption is reasonable because both PRIAM and CatFam are based on the Swiss-Prot database, and their release dates are within 4 months of each other. The resulting distribution of the query proteins in these two datasets against the proteins used for profile generation in CatFam are shown in Figures 2(a,b), respectively, where the MSIs are binned in 5% increments, ranging from 15 to 95%. Figures 2(c–g) show precision and recall results for PRIAM and the 1 and 25% FPR CatFam databases as a function of MSI. The performance of the 5 and 10% FPR CatFam databases is not shown because they have similar trends and are bounded by the presented plots.

We observe similar trends for all precision curves: precision is kept high and does not significantly decrease until the MSI decreases to a turning point, from which precision decreases with decreasing MSI. This point corresponds to 25–30% MSI for the CatFam databases and about 35–40% MSI for PRIAM. In particular, for query proteins with more than 30% MSI, catalytic function predictions with the 1% FPR CatFam database can achieve better than 93.0% precision and as much as 98.0% if it is known [like in Figures 2(c,d)] that the query proteins are enzymes. For the CatFam database with 25% FPR, while still high, precision is reduced to 87.0% and 93.0%, respectively, for the composite dataset

and enzyme datasets. PRIAM's precision is consistently below the corresponding values for the two CatFam databases. The plots for recall shown in Figures 2(f,g) show similar behavior. For D_{enz1}, the 1% FPR CatFam database can achieve better than 89.0% recall when the MSI is larger than 45%, whereas such recall is achieved when the MSI is larger than 35.0% for the database with 25% FPR. For D_{enz2}, the MSI needs to be larger than 50% for the two CatFam databases to achieve better than 89.0% recall. PRIAM's recall for either of the two testing datasets rarely achieves the 89.0% mark.

Predictions for multi-domain and multi-functional enzymes

CatFam was not developed to identify functional domains in protein sequences and reveal their connections with catalytic functions. However, we find that the clustering step in the profile generation process does classify multi-functional enzymes, which quite likely contain multiple domains. For example, the training set of 395 enzymes used to generate profiles for EC 2.7.7.48 (RNA directed RNA polymerase) includes 246 multi-functional enzymes, which are grouped into 47 clusters with different catalytic functions in addition to EC 2.7.7.48. One of such clusters contains 20 enzymes with EC 2.1.1.56, and another contains 22 enzymes with EC 3.1.3.33. Out of the 47 clusters, only three clusters contain enzymes that do not share other EC numbers. This suggests that multi-functional enzymes, sharing a common secondary function, are further grouped into clusters according to their shared secondary functions or, perhaps, different domain structures. In addition, training enzymes with multiple functions are used to generate profiles related to each of their EC numbers. This enables the determination of all EC numbers for a multi-functional enzyme. For example, the 1% FPR CatFam database correctly predicts 569 (91%) out of 622 EC numbers that are assigned to 283 multi-functional enzymes in the testing enzyme dataset D_{enz1}, with only three false predictions.

Analysis of false predictions

We analyze the small number of false predictions made by the 1% FPR CatFam database for enzymes in the D_{enz1} dataset and for nonenzymes in the D_{nz} dataset. For the 7600 enzyme in the D_{enz1} dataset, CatFam incorrectly predicts 58 EC numbers (or 0.76%) for 57 enzymes. Table III lists these 58 predictions along with their true EC annotations based on the Swiss-Prot database (November 2006, release). These correspond to 36 distinct EC numbers out of a total of 856 distinct EC numbers in D_{enz1}.

Our analysis finds that eight out of the 58 EC predictions (highlighted with bold font in the table) that are in fact correct. These include two EC numbers predicted for protein DPOL_HBVAW, whose annotation has been revised in the most recent Swiss-Prot database (February 2008). The other six predicted EC numbers represent catalytic activities that either subsume or are subsumed by the EC numbers assigned by Swiss-Prot. For example, EC 2.4.1.21 (starch synthase using ADP-glucose) predicted for proteins SSG1_HORVU and SSG1_MANES are subsumed by EC 2.4.1.242 (starch synthase using either ADP-glucose or UDP-glucose). In another example, EC 2.7.1.1 (hexokinase) is predicted for protein HXK4_RAT, which subsumes EC 2.7.1.2 (glucokinase). For the other 50 false predictions, we find that 39 of them correspond to EC numbers (and functions) that are very similar to the true annotations. The differences are only at the substrate or cofactor levels, which are usually reflected on the last of the four-digit EC number annotation. Such differences account for 30 out of these 39 false predictions. The other nine false predictions are also associated with substrate-level inferences, although related to multiple EC-digit errors. In eight of such cases, CatFam misidentifies NADH dehydrogenase that acts on quinione (EC 1.6.99.5) for NADH dehydrogenase that acts on ubiquinone (EC 1.6.5.3), and in one case CatFam makes the converse mistake, that is, it predicts EC 1.6.99.5 for EC 1.6.5.3.

It should be noted that CatFam does not make systematic errors for particular EC numbers, in that one EC number is not always predicted for another. This is observed in Table III, which shows that only eight EC numbers (underlined) have error rates higher than 10%. Most of these EC numbers are underrepresented in both the training and the testing datasets, significantly contributing to the higher error rates.

For the 11,349 nonenzymes in the D_{nz} dataset, CatFam incorrectly provides an EC number for 47 (or 0.41%), which are distributed through 27 distinct EC numbers. Table IV lists the six EC numbers that are incorrectly assigned to more than one nonenzyme and the corresponding number of enzymes used to train the related CatFam profiles. Except for EC 2.4.1.129, the number of nonenzymes incorrectly predicted is roughly proportional to the number of training enzymes. This is a consequence of the constraints imposed by the specified FPR during the training process. For a large number of training enzymes, a fixed rate of acceptable false predictions yields a large number of incorrectly predicted nonenzymes. The false prediction related with EC 2.4.1.129 is attributed to the small number of proteins (13) used to construct its profile. In addition, CatFam occasionally predicts EC numbers for close homologs of enzymes that do not possess catalytic activity, as they lack the necessary active sites. For example, CatFam predicts EC 3.2.1.17 for proteins SACA3_HUMAN and SACA_MOUSE that have 50% sequence similarity with the training enzymes. However, these two proteins lack catalytic activity because the required residues at positions 122 (Glu) and 139 (Asp) are not conserved. In another case, CatFam predicts EC

Table III
Analysis of 58 Incorrect EC Predictions for the Testing Enzyme Dataset D_{enz1}

Protein accession[a]	Predicted EC	True EC[b]	Error rate (%)[c]	Catalytic function description[d]
LDH_BOTBR	1.1.1.37	1.1.1.27	9.0	L-lactate dehydrogenase [*Malate dehydrogenase*]
LDH_THEMA	1.1.1.37	1.1.1.27		
GPDA_TRYBB	1.1.1.94	1.1.1.8	4.0	Glycerol-3-phosphate dehydrogenase (NAD(+)) [*Glycerol-3-phosphate dehydrogenase (NAD(P)(+))*]
NUOH1_RHOPB	1.6.5.3	1.6.99.5	6.0	NADH dehydrogenase (quinone)
NUOH_AZOSE	1.6.5.3	1.6.99.5		[*NADH dehydrogenase (ubiquinone)*]
NUOH_BORPA	1.6.5.3	1.6.99.5		
NUOH_RHORT	1.6.5.3	1.6.99.5		
NUOH_THICR	1.6.5.3	1.6.99.5		
NUOI1_RHOS4	1.6.5.3	1.6.99.5		
NUOI_RHORT	1.6.5.3	1.6.99.5		
NUOK_RHOCA	1.6.5.3	1.6.99.5		
NUOG_RICCN	1.6.99.3	1.6.99.5	5.0	NADH dehydrogenase (quinone)
NUOG_RICPR	1.6.99.3	1.6.99.5		[*NADH dehydrogenase*]
NDUS2_RECAM	1.6.99.5	1.6.99.3	5.0	
NU1M_METSE	1.6.99.5	1.6.5.3		
ODP2_ACHLA	2.3.1.61	2.3.1.12	33[e]	Dihydrolipoyllysine-residue acetyltransferase [*Dihydrolipoyllysine-residue succinyltransferase*]
AMY_BACCI	2.4.1.19	3.2.1.1	50	1,4-α-D-glucan glucanohydrolase [*Cyclodextrin glucanotransferase*]
SSG1_HORVU[f]	**2.4.1.21**	**2.4.1.242**		**Starch synthase that uses either UDP- or ADP- glucose [*Starch synthase that uses ADP glucose*]**
SSG1_MANES	**2.4.1.21**	**2.4.1.242**		
APT_YERPE	2.4.2.10	2.4.2.7	9.0	Adenine phosphoribosyltransferase [*Orotate phosphoribosyltransferase*]
OAT_OCEIH	2.6.1.11	2.6.1.13	8.0	Ornithine aminotransferase [*Acetylornithine aminotransferase*]
HXK4_RAT	**2.7.1.1**	**2.7.1.2**		**Glucokinase [*Hexokinase*]**
FER_HUMAN	**2.7.10.1**	**2.7.10.2**	4.0	**Protein-tyrosine kinase [*Protein-tyrosine kinase with an additional transmembrane domain*]**
PPK5_SCHPO	**2.7.12.1**	**2.7.11.1**		**Nonspecific serine/threonine protein kinase [*Dual-specificity kinase for both serine/threonine and tyrosine*]**
KAPB_YEAST	2.7.11.1	2.7.11.11	4.0	cAMP dependent protein kinase
KPCD_CANFA	2.7.11.1	2.7.11.13		Calcium-dependent protein kinase
PLK4_MOUSE	**2.7.11.1**	**2.7.11.21**		**Polo serine/threonine protein kinase, catalyzes same reaction but associates with the spindle pole**
PSK1_SCHPO	2.7.11.1	2.7.11.22		Cyclin-dependent protein kinase
ARGA_PSESM	2.7.2.8	2.3.1.1	6.0	Amino-acid *N*-acetyltransferase [*Acetylglutamate kinase*]
DPOL_HBVAW	**2.7.7.49**	**2.7.7.49**		**RNA-directed DNA polymerase**
DPOL_HBVAW	**3.1.26.4**	**3.1.26.4**		**Ribonuclease**
UBC2_MIMIV	2.7.7.6	6.3.2.19	1.0	Ubiquitin protein ligase [*DNA-directed RNA polymerase*]
TREX2_HUMAN	2.7.7.7	3.1.11.2	1.0	3′-5′ exonuclease [*DNA-directed DNA polymerase*]
MGTA_THENE	3.2.1.1	2.4.1.25	14	4-α-glucanotransferase [α-*amylase*]
GUX6_HUMIN	3.2.1.4	3.2.1.91	12	Exoglucanase [*Endoglucanase*]
GUX_CELFI	3.2.1.4	3.2.1.91		
GUNB_CALSA	3.2.1.8	3.2.1.4	11	Endoglucanase [*Endo-1,4-β-xylanase*]
ATPL_PROMO	3.6.3.14	3.6.3.15	0.4	Sodium ion specific ATP synthase [*ATP synthase*]
ULAD_MYCPN	4.1.1.23	4.1.1.85	3.0	3-dehydro-L-gulonate-6-phosphate decarboxylase [*Orotidine-5′-phosphate decarboxylase*]
TRPF_KLULA	4.1.1.48	5.3.1.24	18	Phosphoribosylanthranilate isomerase [*Indole-3-glycerol-phosphate synthase*]
TRPF_ZYGBA	4.1.1.48	5.3.1.24		
PABB_BACSU	4.1.3.27	2.6.1.85	9.0	Aminodeoxychorismate synthase [*Anthranilate synthase*]
LYS4_SCHPO	4.2.1.33	4.2.1.36	3.0	Homoaconitate hydratase [*3-isopropylmalate dehydratase*]
ISPD_RHOS4	4.6.1.12	2.7.7.60	4.0	4-diphosphocytidyl-2-C-methyl-D-erythritol synthase [*2-C-methyl-D-erythritol-2,4-cyclodiphosphate synthase*]
PHEA_METJA	5.4.99.5	4.2.1.51	50	Prephenate dehydratase [*Chorismate mutase*]
SYWC_BOVIN	6.1.1.1	6.1.1.2	14	Tryptophan translase [*Tyrosine translase*]
SYWC_MOUSE	6.1.1.1	6.1.1.2		
SYWC_RABIT	6.1.1.1	6.1.1.2		
SYW_CLOLO	6.1.1.1	6.1.1.2		
SYN_PYRKO	6.1.1.12	6.1.1.22	4.0	Asparagine translase [*Aspartic acid translase*]
SYT_THET8	6.1.1.15	6.1.1.3	3.0	Threonine translase [*Proline translase*]
SYQ_CLOPE	6.1.1.17	6.1.1.18	5.0	Glutamine translase [*Glutamic acid translase*]
SYQ_PSESM	6.1.1.17	6.1.1.18		
SYK_STRMU	6.1.1.20	6.1.1.6	4.0	Lysine translase [*Phenylalanine translase*]
SYMC_CAEEL	6.1.1.20	6.1.1.10	4.0	Methionine translase [*Phenylalanine translase*]
E2AK4_HUMAN	6.1.1.21	2.7.11.1	3.0	Nonspecific serine/threonine protein kinase [*Histidine translase*]

[a,b]Protein accessions and their true EC numbers are obtained from the Swiss-Prot database released in November 2006.
[c]Error rate is the percentage of false predictions for a given EC number.
[d]Official enzyme names for the true EC number (normal font) and the predicted EC number (italic font in the square brackets).
[e]Error rates greater than 10% are underlined.
[f]EC predictions that are in fact correct are highlighted by bold font.

Table IV
Distribution of EC Numbers Incorrectly Assigned to More Than One Nonenzyme in the D_{nz} Dataset

Predicted EC number	Number of nonenzymes incorrectly predicted	Number of training enzymes
2.4.1.129	2	13
2.4.2.7	2	183
2.7.7.48	4	432
2.7.7.6	13	1762
3.1.1.4	2	270
3.2.1.17	2	155

3.1.1.4 (phospholipase) for protein PA2H_ZHAMA, which has 82% sequence similarity with the training enzymes and has active sites and sequence patterns recorded by PROSITE[34] for that EC number. However, experimental studies do not show catalytic activity for that protein.[35]

Catalytic function annotation for whole genomes

To evaluate the performance of CatFam for whole genome annotation, we select two *Yersinia* genomes [*Y. pestis mediaevails* (ypm) and *Y. pseudotuberculosis* IP 32953 (yps)] and 11 category A and B bacterial pathogens listed by the Centers for Disease Control and Prevention [*Bacillus anthracis* Ames Ancestor (bar), *Burkholderia mallei* ATCC 23344 (bma), *Burkholderia pseudomallei* K96243 (bps), *Brucella melitensis* 16M (bme), *Clostridium botulinum* Hall (cbh), *Coxiella burnetii* RSA 493 (cbu), *Francisella tularensis* SCHU S4 (ftu), *Rickettsia prowazekii* Madrid E (rpr), *Salmonella enterica* serovar Typhi CT18 (sty), *Vibrio cholerae* N16961 (vch), and *Y. pestis* CO92 (ype)]. For benchmarking purposes, we consider the enzyme annotations in the KEGG database (http://www.genome.jp/kegg/) as the gold standard, since these annotations combine the results of multiple resources and are partially curated. Figure 3 shows the CatFam results along with the predictions obtained with PRIAM and EFICAz (http://cssb2.biology.gatech.edu/EFICAz/). In this test, we use the genome-oriented release of PRIAM, which is slightly different from the gene-oriented release used in the tests discussed earlier. EFICAz's predictions and the KEGG's annotations are directly downloaded from their Web-sites in September 2007. Here, we use the CatFam database with FPR preset to 10% because it provides a good trade-off between precision and recall. Figure 3(a) shows the fraction of catalytic function predictions that agrees with KEGG's annotations, that is, it provides a measure of precision, whereas Figure 3(b) shows the fraction of KEGG's catalytic function annotations that are predicted by each method, that is, recall.

Comparison of the three methods indicates that the CatFam predictions yield the largest precision for all 13 genomes [Figure 3(a)]. EFICAz's precision is almost as good as CatFam's, both of which are substantially better than that of PRIAM. All of CatFam's precisions are in the 70–80% range except for *C. botulinum*. This genome was recently sequenced, and only 21 of its proteins are recorded in the most recent Swiss-Prot database. The lack of appropriate training proteins may explain why both CatFam and PRIAM reach their lowest precision values, which are about 55 and 40%, respectively, for this genome. The EFICAz Web-site does not provide predictions for *C. botulinum*. Figure 3(b) shows that CatFam yields the highest recall for seven genomes and that in three cases its recall is substantially lower than that of PRIAM. This is consistent with the fact that often PRIAM predicts many more enzymes than the other two methods, increasing recall at the expense of deteriorating precision. Compared with PRIAM, both CatFam and EFICAz are more conservative tools, optimized for accurate enzyme function predictions.

Despite the overall comparable performance of CatFam and EFICAz, we observe substantial differences when comparing the predictions for each of the 13 bacterial genomes. Similar differences are observed when comparing with PRIAM's predictions as well. For example, the Venn diagram in Figure 4 shows the overlap of the three methods' catalytic function predictions for *Y. pestis* CO92 (*ype*). Although the majority of the three methods'

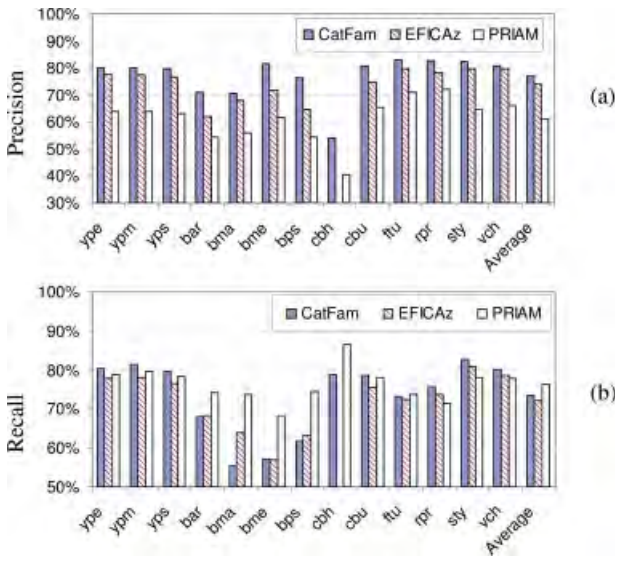

Figure 3
Comparison of catalytic function predictions based on CatFam, EFICAz, and PRIAM for 13 bacterial genomes, using KEGG as the gold standard. (**a**) shows the fraction of catalytic function predictions that agrees with KEGG's annotations, that is, precision. (**b**) Shows the fraction of KEGG's catalytic function annotations that are predicted by each method, that is, recall. The rightmost bars in each of the two panels indicate the average values over the 13 genomes. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

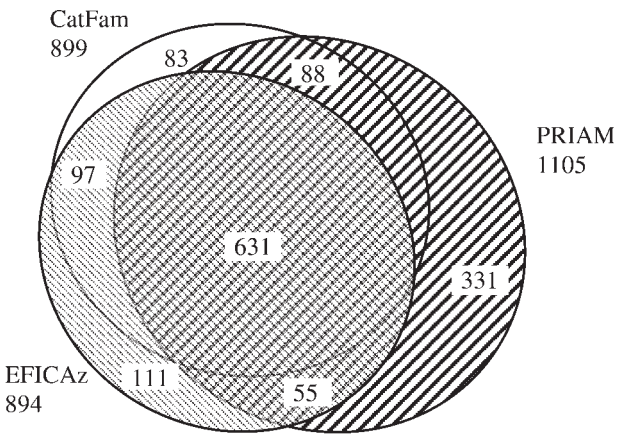

Figure 4

The Venn diagram of catalytic function predictions based on CatFam, EFICAz, and PRIAM for *Y. pestis* CO92 (ype). The number of common predictions is labeled in each intersecting area. The number of predictions solely provided by each method is labeled in each nonintersecting area. The total number of annotations from each method is labeled outside of the diagram.

predictions yield the same functions (631), each provides additional catalytic function predictions that are not inferred by the others. PRIAM provides the largest number of unique predictions (331), consisting of 30.0% of its total predictions. Between CatFam and EFICAz, about 19.0% of the predictions from one method are not provided by the other. We observe similar results for all bacterial genomes except for *C. botulinum*. On average, 21.9% of CatFam's predictions are not inferred by EFICAz and 26.0% of EFICAz's predictions are not inferred by CatFam. False predictions may contribute to some of these unique predictions. However, we expect the majority of these differences to be attributed to slight methodological differences, especially for the differences between CatFam and EFICAz, which are designed for making highly accurate predictions.

Automated metabolic pathway reconstruction

Reconstruction of an organism's metabolic pathways is a key element for understanding the meaning of protein functions within a cellular context. Manual curation is the best way to obtain high-quality metabolic pathways, but it is labor intensive and time consuming. Several tools for automated metabolic pathway reconstructions have been developed (http:/www.pathguide.org). However, the quality of the reconstructed pathways is highly dependent on the precision and extent to which the organism's enzymatic functions are known or predicted. We reconstruct the metabolic pathway of two organisms, *Y. pestis* CO92 and *F. tularensis* SCHU S4, employing the Pathway Tools software.[36] Initially, GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) is used as the sole source of enzyme function annotation and then it is employed in combination with each of the three methods (CatFam, PRIAM, and EFICAz).

Table V compares the total number of reactions and pathways predicted by the PathoLogic-module of Pathway Tools. As expected, the number of predicted reactions and pathways increases as automated enzyme annotations are added to GenBank. The number of predicted pathways is largest when the PRIAM predictions are added to GenBank because, usually, PRIAM provides the largest number of enzyme function predictions for a given genome. However, according to the analysis discussed earlier, some of these predictions are false positives. Biasing the reconstruction toward false positives may be desirable to provide more information to manual curators. For such an application, PRIAM is a valuable tool and is already in use.[37] A combination of multiple prediction tools with complementary enzyme coverage is also valuable, as demonstrated by the joint predictions of CatFam, PRIAM, and EFICAz with GenBank. Combined, when compared against GenBank, they increase the number of predicted pathways for *Y. pestis* and *F. tularensis* by 22.6 and 19.8%, respectively. When manual curation

Table V

Comparison of Predicted Pathways for *Y. pestis* and *F. tularensis*

Organism	Annotation	Number of enzyme-catalyzed reactions	Number of predicted pathways	Number of pathways with holes
Yersinia pestis CO92	GenBank	1000	239	142
	GenBank + CatFam	1060	254	133
	GenBank + PRIAM	1186	278	145
	GenBank + EFICAz	1088	261	141
	GenBank + CatFam + PRIAM + EFICAz	1226	284	145
Francisella tularensis SCHU S4	GenBank	717	169	106
	GenBank + CatFam	745	178	107
	GenBank + PRIAM	818	188	113
	GenBank + EFICAz	754	181	106
	GenBank + CatFam + PRIAM + EFICAz	859	194	114

Various enzyme annotations are used for reconstruction: GenBank, GenBank enhanced by CatFam predictions, GenBank enhanced by PRIAM predictions, GenBank enhanced by EFICAz predictions, and GenBank enhanced by all of the three automated prediction methods.

is not available or not feasible because of the large number of sequenced genomes, more precise prediction tools, such as EFICAz and CatFam, are more appropriate. The number of predicted pathways for these tools is similar, with approximately 95% overlap.

More than 50% of the pathways contain enzymes without assigned genes, the so-called "pathway holes." Some of the pathway holes are filled as automated enzyme annotations are provided by the three methods, but many remain to be filled, perhaps by using the pathway hole filler module of PathoLogic.[38,39] Fully automated application of pathway hole filler introduces additional false positives and is not used for pathway reconstructions in this analysis.

DISCUSSION

The presented results indicate that the CatFam enzyme profiles are effective in discriminating enzymes from nonenzymes, and in predicting a broad range of protein catalytic functions. Although the issue of multi-domain, multi-functional enzymes is not especially considered in the profile generation process, different domain combinations are represented by distinct profiles, enabling the prediction of catalytic functions for multi-functional enzymes.

We observe that four-digit EC numbers do not always classify catalytic functions into distinct categories. A catalytic function classified by one EC number may be subsumed by other function classified by a different EC number. Therefore, we argue that eight of 58 false predictions from a testing set of 7600 enzymes are in fact correct. These include two predicted EC numbers for one protein whose annotation were revised in the most recent Swiss-Prot database, suggesting that CatFam may be robust to under-annotation errors in Swiss-Prot. Conversely, our analysis reveals CatFam's limitations in distinguishing enzymes with very similar catalytic functions. Enzymes that catalyze the same type of reaction but act on different, yet very similar, substrates or require different cofactors are difficult to distinguish and may be missed by CatFam. In addition, CatFam occasionally predicts EC numbers for nonenzymes that are homologous to known enzymes but do not possess active sites. Furthermore, precision control through FPRs may also give rise to a relatively large number of false predictions for EC numbers that are overrepresented in the training dataset.

CONCLUSIONS

We present a new method termed CatFam that generates enzyme sequence profiles to infer protein catalytic functions. The method provides a procedure for specifying the nominal FPR of each profile, thereby controlling the reliability of the predicted protein functions. This enables the generation of profile databases not only for highly precise function annotation but also for moderately precise annotation with better recall, which can be useful for generating hypothetical protein functions. The use of profile-specific thresholds also ensures equal precision for each profile and avoids the problem of having a single *E*-value threshold for all profiles, which yields good overall results but poor performance for some profiles.

Comparisons with well-established resources demonstrate the effectiveness of the enzyme profile generation method and the CatFam databases. They not only achieve overall excellent precision and recall but also perform well for enzymes with low sequence identity. Comparisons based on three testing datasets and 13 bacterial genomes consistently indicate that CatFam outperforms PRIAM in precision and, most of the time, in recall as well. In addition to various improvements in the profile generation procedure, use of negative samples and profile-specific thresholds may be the major contributors for CatFam's superior performance. This is supported by the consistently high precision of CatFam in discriminating enzymes from nonenzymes, whereas PRIAM's performance deteriorates in such applications.

Overall, comparisons between CatFam and EFICAz on whole-genome annotation examples indicate very similar performance. This could be attributed to the similar procedure used to generate the CatFam databases and one of EFICAz's databases. However, the predictions do not completely overlap. On average, 21.9% of the catalytic function predictions inferred by CatFam for 13 bacterial genomes (excluding *C. botulinum*, which EFICAz does not provide predictions for) are not inferred by EFICAz, whereas 26.0% of EFICAz's predictions are not inferred by CatFam. This is perhaps due to methodological or training dataset differences in the profile generation. Although further comparisons may reveal when each method performs best, for now, it seems appropriate to use them complementarily, even combined with PRIAM, for more comprehensive enzyme annotation. We observe a roughly 20% increase in coverage in the reconstruction of metabolic pathways when we combine the predictions of all three methods.

DISCLAIMER

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This article has been approved for public release with unlimited distribution.

REFERENCES

1. Hall N. Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol 2007;210 (Part 9):1518–1525.

2. Pellegrini-Calace M, Soro S, Tramontano A. Revisiting the prediction of protein function at CASP6. FEBS J 2006;273:2977–2983.
3. Medigue C, Moszer I. Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 2007;158:724–736.
4. Friedberg I. Automated protein function prediction—the genomic challenge. Brief Bioinform 2006;7:225–242.
5. Freilich S, Spriggs RV, George RA, Al-Lazikani B, Swindells M, Thornton JM. The complement of enzymatic sets in different species. J Mol Biol 2005;349:745–763.
6. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32 (Database issue):D277–D280.
7. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2004;32 (Database issue):D438–D442.
8. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. Nat Protoc 2007;2:727–738.
9. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;36:307–340.
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
11. George RA, Spriggs RV, Bartlett GJ, Gutteridge A, MacArthur MW, Porter CT, Al-Lazikani B, Thornton JM, Swindells MB. Effective function annotation through catalytic residue conservation. Proc Natl Acad Sci USA 2005;102:12299–12304.
12. Sigrist CJ, De Castro E, Langendijk-Genevaux PS, Le Saux V, Bairoch A, Hulo N. ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics 2005; 21:4060–4066.
13. Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32 (Database issue):D129–D133.
14. Pouliot Y, Karp PD. A survey of orphan enzyme activities. BMC Bioinformatics 2007;8:244.
15. Karp PD. Call for an enzyme genomics initiative. Genome Biol 2004;5:401.
16. Lespinet O, Labedan B. Puzzling over orphan enzymes. Cell Mol Life Sci 2006;63:517–523.
17. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA. Prediction of enzyme classification from protein sequence without the use of sequence similarity. Proc Int Conf Intell Syst Mol Biol 1997;5:92–99.
18. Jensen LJ, Skovgaard M, Brunak S. Prediction of novel archaeal enzymes from sequence-derived features. Protein Sci 2002;11:2894–2898.
19. Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. Proteins 2004;55:66–76.
20. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res 2004;32:6437–6444.
21. Chiu SH, Chen CC, Yuan GF, Lin TH. Association algorithm to mine the rules that govern enzyme definition and to classify protein sequences. BMC Bioinformatics 2006;7:304.
22. Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. J Mol Biol 2005;345:187–199.
23. Levy ED, Ouzounis CA, Gilks WR, Audit B. Probabilistic annotation of protein sequences based on functional classifications. BMC Bioinformatics 2005;6:302.
24. Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA. CORRIE: enzyme sequence annotation with confidence estimates. BMC Bioinformatics 2007;8 (Suppl 4):S3.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
26. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 2003;31: 6633–6639.
27. Tian W, Arakaki AK, Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 2004;32:6226–6239.
28. Arakaki AK, Tian W, Skolnick J. High precision multi-genome scale reannotation of enzyme function by EFICAz. BMC Genomics 2006; 7:315.
29. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
30. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res 2003;31:3497–3500.
31. Artamonova II, Frishman G, Gelfand MS, Frishman D. Mining sequence annotation databanks for association patterns. Bioinformatics 2005;21 (Suppl 3):iii49–iii57.
32. Ouzounis CA, Karp PD. The past, present and future of genome-wide re-annotation. Genome Biol 2002;3;COMMENT2001.
33. Jain AK, Murthy MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999;31:264–323.
34. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 2006;34 (Web server issue):W362–W365.
35. Mebs D, Kuch U, Coronas FI, Batista CV, Gumprecht A, Possani LD. Biochemical and biological activities of the venom of the Chinese pitviper *Zhaoermia mangshanensis*, with the complete amino acid sequence and phylogenetic analysis of a novel Arg49 phospholipase A2 myotoxin. Toxicon 2006;47:797–811.
36. Karp PD, Paley S, Romero P. The pathway tools software. Bioinformatics 2002;18 (Suppl 1):S225–S232.
37. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Medigue C. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res 2006;34:53–65.
38. Green ML, Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics 2004;5:76.
39. Green ML, Karp PD. Using genome-context data to identify specific types of functional associations in pathway/genome databases. Bioinformatics 2007;23:i205–i211.